教学IRでの決定木分析の活用 ： 初年次の学修成果に影響する入学時の学生特徴の探索を例として

著者	紺田 広明, 森 朋子
雑誌名	関西大学高等教育研究
巻	8
ページ	69-78
発行年	2017-03-01
その他のタイトル	Decision Tree for Institutional Research in the Educational Field : As an Example of Searching for Student Characteristics at the Time of Admission Affecting the Learning Outcomes of the First Year Experience
URL	http://hdl.handle.net/10112/11102

教学 IR での決定木分析の活用
―初年次の学修成果に影響する入学時の学生特徴の探索を例として―
Decision Tree for Institutional Research in the Educational Field: As an Example of Searching for Student Characteristics at the Time of Admission Affecting the Learning Outcomes of the First Year Experience

紺田広明（関西大学教育推進部）
森　朋子（関西大学教育推進部）

要旨

本論文では、インスティテューショナル・リサーチ（IR）における主要なデータマイニング手法である決定木分析の活用について論じた。決定木の利点と留意点を整理し、教学 IR における決定木分析の有用性について検討した。具体的な文脈として、初年次の学修成果に影響する入学時の学生特徴を探索するという場面を想定して、決定木の活用の全体の流れを示した。分析設計を検討し、模擬データを用いて R での決定木分析の実行例を提示した。そして、決定木分析が示唆する結果を確認的に分析することで、主要な説明変数を見出すことになり、学習支援等への検討や次なる分析対象の絞り込みにつながることを論じた。

キーワード　教学 IR、決定木分析、データマイニング、高等教育／Institutional Research, Decision Tree, Data mining, Higher Education

１．はじめに

IR（Institutional Research）は、全国の大学でその重要性が認識され始めて、IR 組織の設置が推進されてきている。また、IR の現場では、データベースの構築が進んでおり、学生調査、入試、成績、キャリア関連データなどの多様で大きなデータが集まってきている。一方で、収集したデータを有意味な情報として処理するための主要な道具である統計手法に関しては、取り扱うデータの多様性もあり、まだ模索されている状況である。

データが増えるにつれて、IR の機能の面でも期待されることは大きくなり、積極的な役割を求められるようになってきている。教学を対象とする教学 IR へのニーズの１つとして、潜在する教学上の課題や問題の掘り起こしがある。調査結果の図表化やクロス集計化だけでなく、多様で大きなデータを対象に学習支援やカリキュラム編成等に活かす有用な情報を引き出すことが求められる。これは、データマイニングと呼ばれる考え方である。

主要なデータマイニング手法として、決定木分析（Decision Tree）がある。決定木は、IR 研究における主要なデータマイニング手法として主要な統計手法の 1 つと位置付けられる（Luan, Kumar, Sujitparapitaya, & Bohannon, 2012）。しかし、日本ではあまり使用されていない。木村・西郡・山田（2009）によると、日本での大学生調査に決定木を用いた分析は、山田（2009）が嚆矢とされる。山田（2009）は、決定木により、大学での経験の満足度に影響する変数を探索した。全体的な授業の質の満足、専門分野の授業の内容、学生同士の交流、友人関係、大学への適応といった要因の影響を検討している。また、IR においては、雨森・松田・森（2012）が決定木分析を行っている。3 年次前期の単位修得状況に対して、入

学時の状況（入試形態、本意不本意意識、プレスメントテストなど）、1年次の修学状況（英語、自専攻などの単位数）、1年次に受けた修学サポートなどのうち何が関連しているのかを探索している。その結果から、修学サポートプログラムや科目の改善の提案に結びつけている。このように、主要なデータマイニング手法である決定木であるが、日本の IR での適用はまだ少ない状況である。その手法の特徴や教学 IR での活用に対する有用性の認識は十分とは言えない。

そもそも、統計手法としての決定木の独自性はどこにあるのであろうか。データを分類する類似した統計手法として、クラスター分析（Cluster Analysis）が有名である。クラスター分析では、集団内に似た性質を持つ集まりを見出す。しかし、結果としてのクラスターがどのようなまとまりであるかについては、事後的に特徴を検討することが必要となる（宇佐美, 2017）。また、一般的に使用される回帰分析は、説明変数における加法性を仮定している。仮説が絞れているときは良いが、多数の説明変数があるときや、変数間に複雑な交互作用のある可能性が高いときには、データの傾向を正しく理解できないことがある（Roff, 2006）。詳細は次節において述べるが、これら他の分析手法では実現できない特徴を持っているのが、決定木分析である。

本稿では、教学 IR の多様で大きなデータを対象として、探索的に有意味な情報を見出すというニーズに応えるために、これまであまり報告がなされていない決定木分析の活用への流れに関して検討を行う。以下、決定木に関しては、下川・杉本・後藤, (2013)、Qina (2009)、宇佐美 (2017)、松田・荘島 (2015) などを参考としながらその利点と留意点を整理し、教学 IR の具体的な場面を想定してその適用について論じる。

2．決定木分析の概要

2.1 決定木分析とは

決定木は、端的には、応答変数の反応を、説明変数による単純な識別規則（分枝）を組み合わせて識別していき、反応がより均一な識別領域を得る統計手法である。ここでの識別規則は、応答変数の均一性を高めるように設定されており、逸脱度の最小化や分枝の有意性を検討するものである。結果として得られる識別領域は、リーフとも呼ばれ、説明変数群の基準に従った場合に得られる応答変数の反応をあらわす。

具体的には、図1のように、分析結果は樹状の階層構造で表現できる。

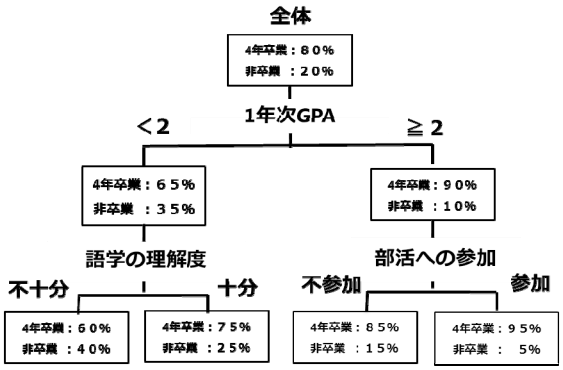

図1 決定木の分析結果のイメージ

この例では、学士課程を 4 年間で卒業できたか、できなかったかを応答変数として、どのような説明変数によってその割合が決まるのかを探索することを課題としている。図において全体と書いている部分は、出発点であり、ルートノードと呼ばれる。応答変数全体の4年卒業率は80%、4年で卒業できない率は20%であることを示している。ここから分枝が行われている。1 つ目の説明変数として選択されたのは「1年次 GPA」であり、「≧2」の場合は右へ、「＜2」の場合は左へと分けられる。これにより、1年次 GPA が 2 以上の者は4年卒業率が 90%、2 未満の者は 65%となる。このように、応答変数の反応を、識別規則に従った説明変数の基準で次々に識別していき、リーフ（ターミナルノード）を得る。この例では、4 つのリーフが得られており、例えば、左下のリーフでは「1 年次 GPA が 2 未満」かつ「語学の理解度が不十分」である場合、4 年卒業率は 60%となる。

また、樹状の階層性は、説明変数の相対的な重要度を示唆する。これにより、4 年で卒業できた者を説明するであろう変数を特定することができる。この例では 4 年卒業率に最も影響しているのは、「1 年次 GPA」であることが示唆されている。

ここでは、応答変数の尺度水準は 2 値の名義尺度（4 年間で卒業できたか、できなかったか）であったが、量的変数を用いる場合もある。それぞれ分類木（カテゴリー変数）、回帰木（量的変数）と呼ばれることがある。また、決定木は、識別規則によって分類される。主要なものとしては、逸脱度度を最小化する方法として、分類回帰木（Classification And Regression Tree：CART）、分枝の有意性を検討する方法は、カイ２乗自動交互作用検出（Chi-squared Automatic Interaction Detector: CHAID）がある。

分類回帰木を例として、識別規則を説明すると、次のような考えに基づいている。分枝がうまく行われたならば、各ノード内に含まれる応答変数の反応は均一になる。つまり、応答変数の分散が小さくなることが良い分枝であると考えられる。この考え方から、分枝には応答変数の偏差平方和である逸脱度という指標が用いられる。回帰木（反応変数が量的変数）において、逸脱度は、$D_i = \sum_{k=1}^{m_i}(y_k - \mu_i)^2$とされる。ここで、$D_i$は、観測数$m_i$をもつi番目のノードの逸脱度であり、$y_k$はそのノードのk番目の観測値、そして$\mu_i$はノードiの予測される平均値である。分類木（反応変数がカテゴリー変数）の場合は、情報量指数やジニ係数で定義される。逸脱度を用いて、分枝前と後での差を比較して最も低下する分枝点を有する変数が、最も良い説明変数として選択される。また、カイ２乗自動交互作用検出では、分枝後において分枝基準でのカイ２乗検定が有意であることを基準としている。これは、有意味な分枝点を見出そうとしていて少し考え方が異なる。

上記の識別規則は、分枝点でのデータを対象にして、繰り返し行われることで木は成長する。これは、データへの過適合を起こしうる。そこで、予測精度などを参考にして適切に刈り込みを行うことで、分析結果を得ることになる。

2.2 決定木分析の利点と留意点

決定木の主な利点としては、以下の点が挙げられる（下川・杉本・後藤, 2013; Qina, 2009）。

① 結果をグラフィカルに提示することができ、解釈が比較的容易
② どの説明変数を使用するかを決める必要がない（重要な説明変数が自動的に選択される）
③ 交互作用や非線形構造を自然に扱える
④ 説明変数は、データ数よりも大きくてもよく、カテゴリー変数でも量的変数でも適用できる

①に関しては、先ほど図 1 を用いて説明した通りである。教学 IR での使用を考えた場合、グラフィカルで簡単なルールとして説明できる分析結果であることは特に重要な利点であろう。②は、決定木は影響しそうな説明変数が非常に多くある場合に有効である。教学 IR の現場では、仮説を明確に絞ることができず、説明変数を特定できていない場面は少なくない。先ほどの例でも 4 年卒業率に影響しそうな要因は無数にあるだろう。決定木は、それらを識別規則に基づいて、言わば自動的に、応答変数の反応を均一にする説明変数を選び出す。その際、階層性の上部で選ばれる変数は、応答変数を予測するために重要な説明変数として特定することができる。③に関しては、説明変数の分枝を繰り返していくため、応答変数を区分するルールの中に、説明変数間の影響関係を自然に組み込んでいることになる。先ほどの例においても、左下のリーフでは「1 年次 GPA が 2 未満」かつ「語学の理解度が不十分」である場合、4 年卒業率は 60%であるとなり、説明変数の組み合わせが明らかになる。④に関しては、説明変数において特段の処理や注意を払わなくてもよいということである。GPA のような量的変数でも使用可能であり、併願状況（国公立併願、私大併願、

本学専願）などのカテゴリー変数でも同時に適用できる。説明変数の数においても、多数の候補となる変数を同時に投入することが可能である。

一方で、決定木分析の欠点や留意点としては、次のようなことが挙げられている（下川・杉本・後藤, 2013; Qina, 2009）。分析で特定された少数の説明変数のみで予測することになるため、予測精度は一般に高くないとされる。つまり、応答変数を予測することを目的として、決定木を使用することはあまり有用ではない。予測精度を上げるためには、決定木を多数発生させて統合するランダムフォレスト（RandamForest）法といった手法が用いられる。ただし、一般に結果の解釈が困難になる。

また、別の留意点として、データセットの違いや識別規則により、異なったモデルとなりうる。決定木による結果はわかりやすい応答変数を区分するルールをもたらすが、識別規則を変えることで、結果は変わりうる。時として、単純にリーフに残る個数を変更しても結果は変わることがある。対応策としては、識別規則などを変えて適用して検討することが必要となる。このことと関わるが、決定木は、分枝ごとで最適な分枝を作成していて、モデル全体を考慮して分枝しているわけではない。そのため、分枝した中でさらに分枝していくため、分枝された集団を前提として次の説明変数が選ばれている。先に述べたとおり、これは説明変数同士の交互作用を自然な形で取り入れていると考えられるが、分枝に用いられた説明変数に過度に固執するのではなく、代替となる変数はないかなどの検討が必要となる。加えて、応答変数と説明変数の関係において線形構造をもつとき、決定木の適合性能は急激に低下する。これは、分枝は、階段状に応答変数を区切っていることと類似しているためである。

以上まとめると、決定木の主要な利点は、解釈の比較的容易なグラフィカルな結果を提示して、重要な変数を特定することができることである。また、欠点としては、不安定性にあると言えるだろう。これは、決定木が探索的な分析手法であることと表裏一体と言える特徴である。教学 IR では、教務、成績、学生調査、キャリアなどのデータを収集して分析するが、これらの多様で大きなデータから役立つかもしれない情報の可能性を抽出する（データマイニング）手法であると決定木を捉えることが有用であろう。

3．教学 IR での決定木分析

3.1 初年次の学修成果に影響する入学時の学生特徴の探索

本節では、決定木分析によってどのようなことを明らかにできるのかについて検討するため、教学 IR における具体的な文脈のなかで、決定木の適用例を考えてみたい。

ここでは、初年次教育の学修成果に影響を与える要因を探るという課題を扱いたい。問いの形で表すと、「初年次教育の学修成果の違いは、どのような入学時の学生特徴から生じるのか？」であろう。これは、入学時の学生の特徴から初年次教育の学修成果がわかれば、入学初期に学習支援や対応を考えていくことが可能となり、教学の質保証のための PDCA サイクルを回すことに資すると考えられる。

この問いに答えるためには、まず初年次教育の学修成果は、どのような指標で表現できるのかについて検討が必要になる。初年次教育の成否を表す指標とは何かについて具体的に考えることとなる。これは、分析においては応答変数（従属変数）にあたる。また、入学時の学生の特徴はどのようにデータとして捉えることができるのかを検討することから始まるであろう。それは、例えば、一般入試や推薦入試などの入試種別として捉えることができるであろう。また、入学時の学生調査の変数を使い表現すること、あるいは不本意入学などの入学時の学生が抱える課題や状況について考えることが必要になる場合もあるだろう。これは、分析における説明変数（独立変数）の構成に関わることである。

また、この問いの困難さの１つは、入学時点での学生の特徴と考えられるものは非常に多くあることである。ミスマッチ、入学時不安、コンピテンシー、パーソナリティ、将来の進路希望、高校での正課外活動、・・・。学生の特徴を捉える視点は多様にありえる。初年次教育に関して、効果的な授業プログラムや指導内容に関する研究はあると考えられるが、入学時の学生の特徴が初年次教育の成否に及ぼす影響についてはわからない点が多い。つまり、仮説を明確には絞れない。

仮説が絞れていれば、回帰分析を行うのが一般的であろう。回帰分析は、１）特定の説明変数（群）が応答変数へ与える影響を検討、あるいは２）特定の説明変数（群）で応答変数を予測、を主な目的として使用される。例えば、「入学時に不安を抱える学生は、GPA が低い」との特定の仮説があれば、回帰分析（説明変数：入学時不安、応答変数：GPA）して、入学時不安の係数が、負で有意であれば、仮説通りとわかる。つまり、回帰分析は、一般にはどの説明変数を使用するのかを事前に知っていなければならない。

今回は、説明変数（入学時の学生調査等から見える入学時点での学生の特徴）が非常に多くあって、応答変数（入学時教育の学修成果）にどれが影響を及ぼすかの仮説が明確にはわからない。そのため、仮説生成的に決定木分析を用いることになる。課題の概要を図として示すと以下になる。

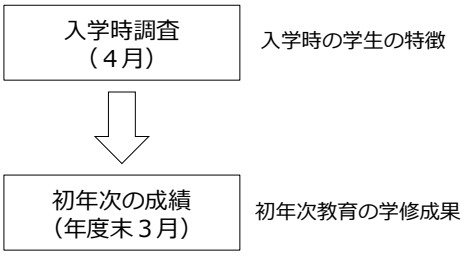

図２　入学時の学生特徴からの
初年次の学修成果の予測

3.2 R での決定木分析

決定木分析は、SPSS などでも解析が可能であるが、ここでは、R の rpart パッケージの関数を用いて分類回帰木（CART）で分析してみる。なお、R での決定木の詳細に関して、下川・杉本・後藤,（2013）、Qina（2009）、外山・辻谷（2015）が詳しい。

分析データとしては、模擬的なデータを準備した。100 名を対象として、初年次教育の学修成果の指標として成績を想定し、応答変数は「1 年次 GPA」とした。この値は正規乱数（平均 3、標準偏差 0.7）として発生させた。これを説明する変数として、「大学への目的」「入試形態」「入学時不安」を想定した質的変数を 3 つ、「高校時成績」「入学前の課題成績」を想定した量的変数 2 つを用意した。ここでは説明の分かりやすさのために、この 5 つの変数は任意に値の割り振りをして「1 年次 GPA」に関連すると考える変数として設定した。また、説明変数が非常に多くある状況として、「1 年次 GPA」と関連しないと想定した正規乱数に従う 70 個の変数も同時に説明変数として用意した。結果として、これらを 100 行×76 列のデータ d として R に準備した。

決定木の分析を行うには、最初にパッケージ rpart をインストールして、library (rpart)とパッケージの読み込みを行う。その後、R での決定木分析は、基本的には rpart 関数による次のコードである。

```
result <- rpart ( GPA ~ . , data = d, control =
                  rpart.control ( minsplit = 4, cp = 0.001 ) )
```

rpart 関数による結果は、result と名づけた変数に格納している。rpart 関数の括弧内は、引数の指定であるが、応答変数として GPA を指定して、チルダ「~」の後ろに説明変数を指定している。ここでは、データ d に含まれる GPA 以外の変数全てということでピリオド「.」としている。これらは R での回帰式を指定する方法と同様である。

対象とするデータの指定は、data において行っている。また、control 以下は決定木分析の指定であり、minsplit は、各ノードにおける最小の個数（人数）の指定である。ここでは 4 とした。cp は複雑性パラメータと呼ばれる。指定した cp 値によってモデルの複雑さは制限を受けて、木の大きさと関連する。cp 値が小さいほどより複雑なモデルとなる。

最初は、小さい値 0.001 などを指定して、木を成長させた結果を出す。次に、データへの過適合を防ぐための木の剪定作業を行う。printcp 関数を用いて、printcp (result)と入力すると cp 表と呼ばれる結果が示される。これは、cp 値、分枝数（nsplit）、相対誤差（rel error）、交差妥当化誤差（xerror）、その標準誤差（xstd）の表が示される。また、plotcp (result)と入力すると、その表を図とした次のような出力が得られる。

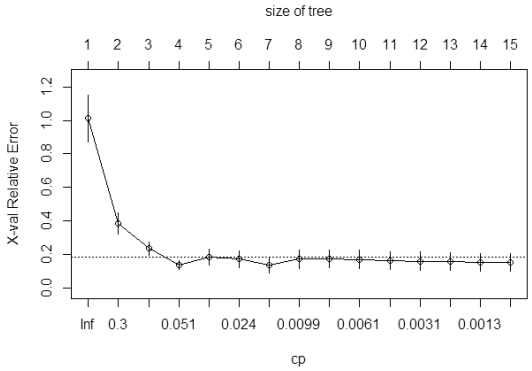

図 3 木の剪定を検討するための cp プロット

この図は、縦軸が、予測精度を示す交差妥当化誤差（xerror）であり、横軸の上に木のリーフの数、それに対応する cp の値が横軸の下に表示されたものである。右に行くほど cp 値は小さくなり、木のリーフの数が増えて複雑になっている。この図と表を検討して、予測精度の観点では、4 つのリーフがあるところが最も小さくなっている。

交差妥当化誤差は小さいほど予測精度が高いことを示すが、同程度であれば分枝の少ない木を選ぶことが行われる。これは、1 標準誤差ルールと呼ばれる方法である（Breiman, Friedman, Olshen, & Stone、1984）。最小の交差妥当化誤差とその標準誤差を足した範囲内で最も小さい木を選ぶ方法である。ここでは、最小の交差妥当化誤差である 4 つのリーフにおいて、交差妥当化誤差（xerror）±1 標準誤差（xstd）の範囲内（図において点線で示されているところより低い）での最も小さな木となるが、点線より低い値を示す 4 のリーフより小さい木はないので 4 のリーフを持つ木を選択することになる。

この予測精度の観点での検討により木の大きさが決まったので、枝の刈り込みを行う。刈り込みは cp 値の指定する（複雑さを制限したモデルを推定する）ことで行うことができる。リーフ 4 での cp 値とリーフ 3 での cp 値との間の適当な値にして指定した以下のコードを実行することで、リーフ 4 での結果が得られる。

```
result2 <- rpart ( GPA ~ . , data = d, control =
                rpart.control ( minsplit = 4, cp = 0.052 ) )
```

結果の樹状図を得るには、plot 関数で可能であるが、より綺麗な図を得るためには、パッケージ partykit をインストールして、library (partykit) しておいて次のように入力する。

```
plot ( as.party ( result2 ) )
```

これにより、図 4 を得ることができる。

この模擬データでの決定木の分析結果では、75 の変数のうち、3 つの変数「大学での目的」、「入学時不安」、「高校時成績」が説明変数としてモデルに使用された。最初の分枝は「大学での目的」によるものであり、資格取得または知識教養は左へ、専門分野は右へと分けられている。次に、左側の分枝では、「入学時不安」によって、不安なしは左へ、不安またはやや不安は右へとなっている。右側の分枝では、「高校時成績」によって、3.583 未満は左へ、3.583 以上は右へと向かっている。

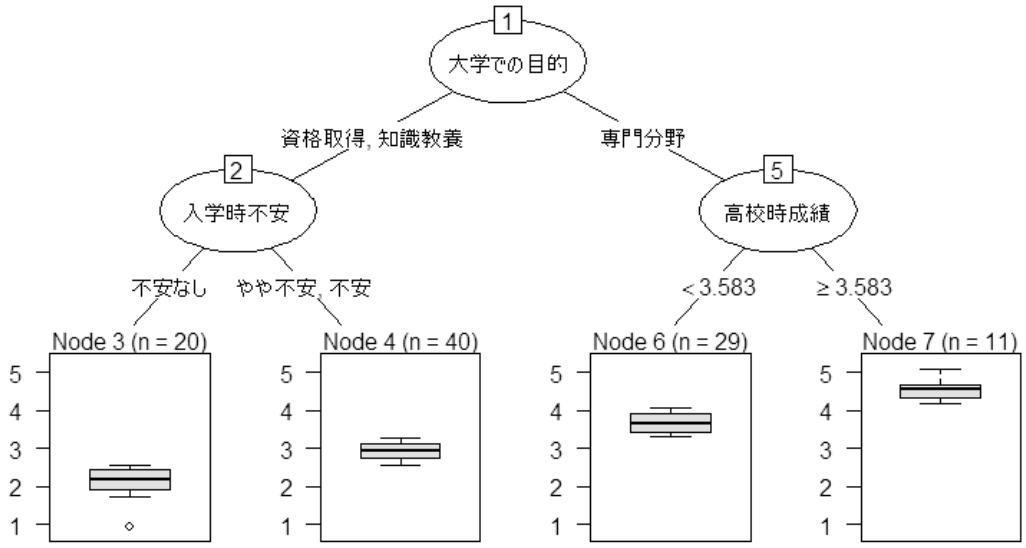

図 4 学修成果に関わる入学時特徴についての決定木分析結果（模擬データ）

最終的には、4 つのリーフが得られていて、それぞれは GPA が最も低い群から順に GPA 低群 20 名、GPA 中低群 40 名、GPA 中高群 29 名、GPA 高群 11 名のグループとなった。

この結果では、正規乱数に基づく 70 の変数は選択に用いられず、応答変数を区分けするための説明変数の最適な組み合わせを見出している。本分析では、これらの変数は任意に設定してはいるが、応答変数を分類するために、多数の候補となる変数からの選択と組み合わせを見出せることは、決定木ならではの強みである。一方で、主要に関連すると設定したものの 2 つの変数は使用されなかった。このように、決定木分析では、多少の関係があったとしても、その分枝ごとにおいて最適ではない変数はモデルから除外されることになる。これは先に述べた通り、決定木の使用上の留意点として注意しなければならない。

3.3 分析結果の確認的分析と活用へ

決定木分析は、探索的なデータマイニング手法であるので、結果から示唆された説明変数がどのような影響を応答変数に及ぼしているのかを、別の角度からも検討することで、より結果が鮮明になる。

決定木の結果、重要であろう 3 つの変数「大学での目的」、「入学時不安」、「高校時成績」が示唆された。つまり、これら入学時での 3 つの要因が、初年次の学修成果と関連するという考えがもたらされた。この結果にもとづき、さらに確認的に分析を続けていく。各分枝での「1 年次 GPA」はどのように異なるであろうか。結果を図 5 として示す。確かに「大学での目的」や「入学時不安」の分枝基準によって GPA は大きく異なることが示される。一方で、「高校時成績」の 3.583 は、全体においては意味がある基準ではないことがわかる。次に、「高校時成績」の影響を検討するために、図 6 を示す。これは、縦軸に 1 年次 GPA、横軸に「高校時成績」、各点の形を「大学での目的」とした散布図である。この図において、垂直な点線は「高校時成績」の 3.583 を示している。先ほど確認した全体で見れば「高校時成績」の基準が GPA の高低を識別できているわけではないのは、その点線の右側には 2 つの集団があるためのようである。しかし、「大学での目的」で専門分野と答えている学生は確かに GPA 値が高く、また、その専門分野のなかでは、「高校時成績」の 3.583 は、より GPA が高いクループを識別しているようである（GPA が 3.5 を超える図における円のな

かでの点線)。また、応答変数を1年次GPAとして、3つの変数を説明変数とした重回帰分析を行った結果、いずれの変数も有意な結果となった。専門分野や高校時成績は正の影響、不安なしは負の影響をGPAに与えていていることがわかった。

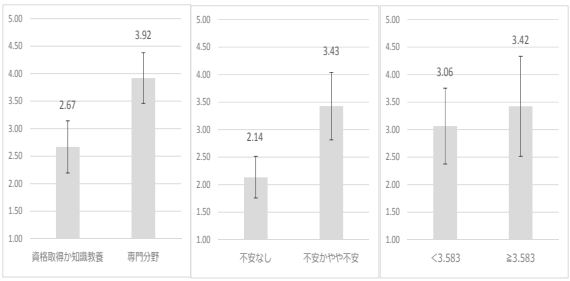

図5 分枝点での1年次GPAの違い

注：左「大学での目的」、中央「入学時不安」、
左：「高校時成績」

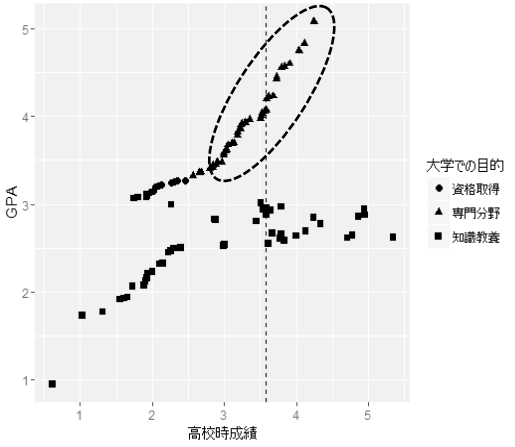

図6 GPAと高校時成績の散布図

本分析の結果をまとめると次のようなことが言える。ここでの教学IRとしての問いは、初年次の学修成果に影響する入学時の学生特徴の探索であった。この問いへの回答は、1つには学生が意識している大学での目的が影響している可能性がある。学問としての専門分野を目的として入学してきている学生は、成績が高くなる傾向がある。また、入学時の不安が生じている学生のほうが、成績は高い傾向を示した。さらに、より成績が高い学生は、高校時の成績が影響していることがわかった。他の要因はこれらの要因と比較して成績には影響を及ぼしてはいない。

この結果をもとに初年次での対応策を考える。雨森・松田・森（2012）で主張されるように、IRの役割は、PDCAサイクルのCheckで終わらせるのではなく、Cにもとづく改善案や次の分析ターゲットを提示することが重要である。潜在的なニーズを教学IRが探索することの意味はここにあるだろう。

入学時不安の結果に関して言えば、不安は一般にネガティブに捉えられが、他方その対象に対する関与の大きさや重要性の認識としても考えられる。大学生活を充実した良い形で過ごしていきたいなど、大学に対して真剣に考えている可能性がある。例えば、上位年次のロールモデルの提示など学生が大学への関与を高めるアプローチを考えていくことができるであろう。

さらに、本分析では、入学時の学生特徴による予測を行っている。これは、入学後の正課（外）での活動についての情報がない中で予測している。授業等で生じる効果も検討することで対応策はさらに明確になるだろう（図7）。

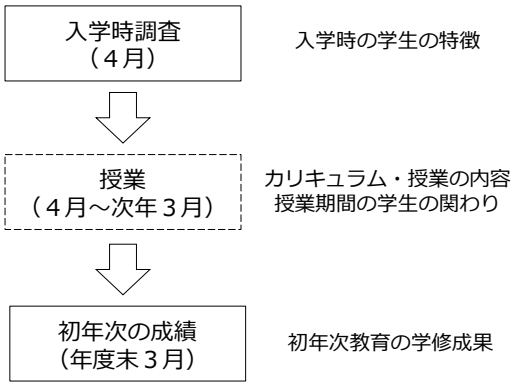

図7 入学時の学生特徴と授業からの初年次の学修成果の予測

例えば、初年次においては、学問としての専門分野に意識を持たせるために、専門入門の授業や、少人数のゼミ形式での専門導入の授業がある。これらにおいて、学生の専門分野への意識、大学での目的がどのように変化したのかを検討すること

も行えるだろ。分析結果は、教育改善の方策を考える糸口に過ぎないが、次のステップへと検討を進めることができる。

4．まとめ

本稿では、教学 IR の多様で大きなデータを対象として、探索的に有意味な情報を見出すというニーズに応えるために、主要なデータマイニング手法である決定木分析の活用に関して検討を行った。決定木分析の利点と留意点を認識した上で、教学 IR における具体的な文脈のなかで、決定木の適用例を考えてみた。ここでは、初年次教育の学修成果に影響を与える要因を探るという課題を R での分析手順を示した。そして、決定木分析が示唆する結果を、確認的に分析することで、主要な説明変数を見出すことになり、学習支援等への活用や次なる分析ターゲットの絞り込みにつながることを論じた。

教学 IR は、学内外にあるデータが集積されるにつれて、データを整理して示す役割だけではなく、教育や学びに活かすために積極的な役割を求められることも多い。その際に、決定木分析の活用の可能性があるだろう。例えば、入学から卒業までに履修する多数の科目において、学生がつまずくキー科目を探索することなどにも使用できる。見出された科目に関して検討を行うことで、学習支援やカリキュラム再編等への足掛かりとすることができるであろう。

一方で、教学 IR の文脈を明確に捉えておくことも必要である。本分析では、GPA を対象にしていたが、初年次教育の成否の指標は、様々に模索されている。例えば、北海道大学の理系実験にける初年次教育では、レポート平均点、提出状況、遅刻回数において、入試制度による違いを検討している（田島, 2014）。また、京都産業大学では、初年次教育の指標としてエニグラム（自我状態を場面に応じて適切に使う力）の有効性を検討している（後藤, 2012）。そして、山田（2008）は、「初年次教育の評価の方法についても、何をもって学生が入学時と比較して獲得したかを測定することの困難さや点数ベースの評価法が適切であるか等課題は大きい（山田, 2008, p.23）」としている。このような教学の文脈を俯瞰した分析視点を持ち、決定木分析を有効に活用していくことが重要である。

参考文献

雨森 聡・松田 岳士・森 朋子 (2012). 教学 IR の一方略：島根大学の事例を用いて　京都大学高等教育研究, *18*, 1-10.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Pacific Grove, CA: Wadsworth.

後藤 文彦 (2012). 初年次教育の有効性に関する実証的研究　高等教育フォーラム, *2*, 1-7.

木村 拓也・西郡 大・山田 礼子 (2009). 高大接続情報を踏まえた「大学教育効果」の測定――潜在クラス分析を用いた追跡調査モデルの提案――　高等教育研究, *12*, 189-214.

Luan, J., Kumar, T., Sujitparapitaya, S., & Bohannon, T. (2012). Exploring and mining data. In R. D. Howard, G. W. McLaughlin, W. E. Knight, & Associates (Eds.) *The handbook of institutional research* (pp.478-501). San Francisco: Jossey-Bass.

松田 いづみ・荘島 宏二郎 (2015). 犯罪心理学のための統計学――犯人のココロをさぐる――誠信書房

Qina, S. S. (2009). Environmental and ecological statistics with R. FL: Boca Raton, CRC Press. (大森 浩二・井上 幹生・畑 啓生 (監訳)(2011). 環境科学と生態学のための R 統計　共立出版)

Roff, D. A. (2006). Introduction to computer-intensive methods of data analysis in biology. Cambridge University Press. (野間口眞太郎 (2011). 生物学のため

の計算統計学――最尤法、ブートストラップ、無作為化法　共立出版)
下川 敏雄・杉本 知之・後藤 昌司 (2013). 樹木構造接近法　共立出版
外山 信夫・辻谷 将明 (2015). 実践 R 統計分析　オーム社
田島 貴裕 (2014). 初年次教育の理系実験に対する取組姿勢と成績評価の関連性――入試制度の観点から――　日本教育工学会論文誌, *38*, 1-4.
宇佐美 慧 (2017). 縦断データの分類――決定木および構造方程式モデル決定木――　荘島宏二郎(編)　計量パーソナリティ心理学 pp.219-239　ナカニシヤ出版.
山田 礼子 (2008). 初年次教育の歴史と理論　大学と学生, *54*, 16-22.
山田 礼子 (2009).学生の情緒的側面の充実と教育成果―CSS と JCSS 結果分析から―　大学論集, *40*, 181-198.

紺田広明（関西大学教育推進部）
森朋子（関西大学教育推進部）